Treatment Recommendation with Preference-based Reinforcement Learning

Nan Xu, Nitin Kamra, Yan Liu
Department of Computer Science, University of Southern California
{nanx, nkamra, yanliu.cs}@usc.edu

***Abstract*—Treatment recommendation is a complex multi-faceted problem with many *treatment goals* considered by clinicians and patients, e.g., optimizing the survival rate, mitigating negative impacts, reducing financial expenses, avoiding over-treatment, etc. Recently, deep reinforcement learning (RL) approaches have gained popularity for treatment recommendation. In this paper, we investigate preference-based reinforcement learning approaches for treatment recommendation, where the reward function is itself learned based on *treatment goals*, without requiring either expert demonstrations in advance or human involvement during policy learning. We first present an open simulation platform [1] to model the evolution of two diseases, namely *Cancer* and *Sepsis*, and individuals' reactions to the received treatment. Secondly, we systematically examine preference-based RL for treatment recommendation via simulated experiments and observe high *utility* in the learned policy in terms of high survival rate and low side effects, with inferred rewards highly *correlated* to treatment goals. We further explore the *transferability* of inferred reward functions and guidelines for *agent design* to provide insights in achieving the right trade-off among various human objectives with preference-based RL approaches for treatment recommendation in the real world.**

I. Introduction

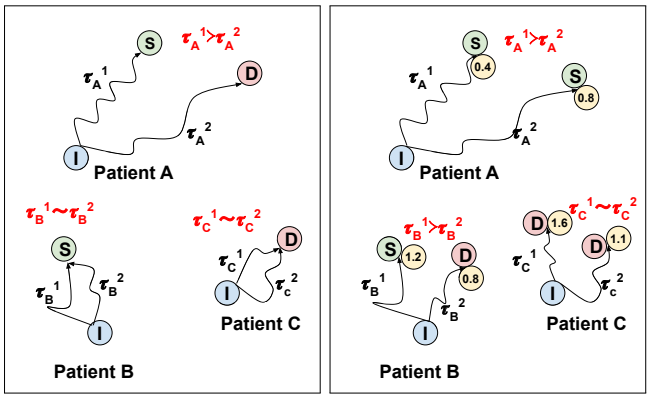
With recent advances in deep learning, deep reinforcement learning (RL) approaches have gained popularity for treatment recommendation [1]–[6]. But the success of RL applications often crucially depends on the prior knowledge that goes into the definition of the reward function [1]–[3], [7]. However, treatment recommendation is a multi-faceted problem where the reward function is hard to engineer and requires quantifying the trade-off among diverse types of *treatment goals*: 1) ***Clinicians*** aim to optimize the survival rate (or expected lifetime) while mitigating negative impacts of treatments [1], [2], [5]; 2) ***Patients*** often consider reducing financial expenses and time costs in accepting treatment strategies [4], [8]; 3) ***Other factors*** are considered to avoid over-treatment or follow agreements based on the medical insurance plan for an affordable treatment [3].

To explicitly reflect humans' *treatment goals* in the reward functions, prior work has jointly considered multiple objectives linearly weighted to reduce the problem to a single-objective MDP [4]–[6], [8]. However, the linearly weighted reward function induces negative interference between objectives (see Section IV-D) and ends with treatment outcomes against humans' actual intentions, especially when representations are learned using neural networks and shared among different objectives [9], [10]. Another line of approach is to infer the clinicians' intrinsic reward function using existing inverse reinforcement learning (IRL) methods [11], [12]. However, these assume that the historical demonstrations comprising of clinicians' treatment strategies are from an optimal policy, which is not guaranteed in practice [13]–[16].

[1]https://sites.google.com/view/tr-with-prl/

Fig. 1: Illustration of preferences over two treatment trajectories given a patient with *Cancer*, based on two *treatment goals*. Starting from an initial state I, the patient adopts a treatment strategy with trajectory τ and ends with either survival S or death D outcome. The final tumor size value is also shown for *Goal II*. A trajectory can be preferred to the other ($\succ$) or the two can be incomparable ($\sim$).

(a) *Goal I*: Survive (b) *Goal II*: Survive with small tumor



Fortunately, while it is hard to obtain demonstrations from an optimal policy, qualitative feedback from clinicians' preferences can be easily obtained and efficiently leveraged to infer reward functions. In this paper, we investigate preference-based reinforcement learning approaches [16]–[23] for treatment recommendation, where a reward function is learned using *treatment goal*-directed preferences over pairs of trajectories. Different from existing work where the preferences mainly come from human annotators [21], [22] or the level of noise applied to suboptimal demonstrations [23], we consider the *treatment goal* itself as the criterion of preference over two treatment trajectories for any sampled subject, without requiring demonstrations in advance or any expert involvement during policy learning. Take *Patient A* in Fig. 1a as an example: when the *treatment goal* is to save the patient's life alone, then the treatment trajectory τ_A^1, which leads to

the survival outcome, is preferred to trajectory τ_A^2 that ends with death. Given preference over two trajectories, we follow the Bradley–Terry model [24] and assume that the preference probability of one trajectory depends exponentially on the reward sum, after which the reward learning problem is turned into a binary classification task.

However, acceptance of preference-based RL approaches for treatment recommendation requires significant exploration in the inferred rewards and achieved performance. This paper investigates the above aspects with the following contributions:

- We first develop an open simulation platform to model the dynamic state transitions of individuals with *Cancer* or *Sepsis*, as well as their reactions to the received medication or operation treatment. The developed simulation platform enables efficient model training and reliable performance evaluation.
- Next, we conduct comprehensive simulated experiments to obtain the following conclusions: 1) *Utility*: The preference-based RL framework outperforms other treatment recommendation baselines with high survival rate and low side effects under different *treatment goals*. 2) *Goal-Correlated*: The reward function inferred from preference is highly correlated to several important factors in *treatment goals*.
- We further explore the *transferability* of the inferred rewards to different scenarios, which shed light to transfer preference-based RL models trained in simulation to the real world with low costs.
- We provide guidelines for *agent design* in preference-based RL framework, i.e., competitive RL agents with diverse policies achieve better treatment outcomes than others.

II. Problem Definition

We cast the treatment policy learning as a Markov Decision Process (MDP)[2], where an agent interacts with an environment over a sequence of steps: at each time step t, the agent receives a state $s_t \in \mathcal{S}$ from the environment and responds with an action $a_t \in \mathcal{A}$. The environment state transition is controlled by the probability function $\mathcal{P}(s_{t+1}|s_t, a_t)$ and the agent continues to interact until a terminal state is reached at time T[3]. Given an initial state s_0, the agent follows a policy π and generates a *trajectory* of sequence $\tau = ((s_0, a_0), (s_1, a_1) \ldots, (s_{T-1}, a_{T-1})) \in (\mathcal{S} \times \mathcal{A})^T$. In traditional reinforcement learning, the agent receives a reward $r_t \in \mathbb{R}$ at each time step and aims to maximize the expected return by summing up the rewards with a discount factor.

Instead of assuming a reward signal from the environment, we consider that a clinician or a patient establishes a treatment goal in advance[4], and preference between two treatment trajectories for a specific patient can be produced naturally accordingly to their goal fulfillment. We use $pre(\tau^1, \tau^2) = \tau^1 \succ \tau^2$ to indicate that trajectory τ^1 is preferred to trajectory τ^2, and $pre(\tau^1, \tau^2) = \tau^1 \sim \tau^2$ for incomparable trajectories.

In Fig 1, we demonstrate an example of dosage recommendation for *Cancer* patients with two types of goals:

- ***Goal I-Survive***: In Fig. 1a, the trajectory leading to survival outcome is preferred to that with death outcome: $pre(\tau_A^1, \tau_A^2) = \tau_A^1 \succ \tau_A^2$; two trajectories are incomparable when they both have identical outcomes: $pre(\tau_B^1, \tau_B^2) = \tau_B^1 \sim \tau_B^2$, and $pre(\tau_C^1, \tau_C^2) = \tau_C^1 \sim \tau_C^2$.
- ***Goal II-Survive with small tumor***: In Fig 1b, considering two trajectories with both survival outcomes, the one resulting in smaller tumor size is preferred: $pre(\tau_A^1, \tau_A^2) = \tau_A^1 \succ \tau_A^2$; trajectories that enable the patient to survive are always preferred over those leading to deaths: $pre(\tau_B^1, \tau_B^2) = \tau_B^1 \succ \tau_B^2$; if neither trajectory results in a survival outcome, they are incomparable: $pre(\tau_C^1, \tau_C^2) = \tau_C^1 \sim \tau_C^2$.

Informally, the goal of the agent is to recommend treatment strategies which are preferred based on humans' *treatment goals*. To achieve this, the agent aims to maximize the expected return with rewards inferred from the preferences, which is explained in detail in the following section.

III. Method

After setting the *treatment goal*, we demonstrate the joint learning framework for reward and policy in Algorithm 1. Given one sampled subject, two agents with their policies parameterized by θ_P^1 and θ_P^2 provide the pairwise trajectories to compare, and a reward model parameterized by θ_R estimates a reward function with preference.

In the beginning, parameters for the reward and policy model are randomly initialized (line 1), and the preference and trajectory samples for the model update are created as empty lists (line 2). In each training iteration, one subject is sampled from the training set to receive treatment from both agents (line 3 to 6). At each time step, the agent makes the treatment decision based on the current state, the simulator updates the subject's status, and the reward model generates the corresponding step-wise reward (line 7 to 12). At the end of the trajectory, the trajectory list is augmented with the latest treatment trajectory (line 13), while the preference list for the reward model is also updated (line 16) with the preference over trajectories from two agents according to the *treatment goal* (line 15). After traversing over all training subjects, a minibatch of the preference samples is extracted to fit the reward function (line 19), while a minibatch of the trajectory samples is utilized to optimize the policies for both agents (line 20).

[2]For simplicity, here we assume all the system variables can be observed directly. In our experiments, policy learning for *Cancer* is MDP while *Sepsis* is Partially Observable MDP, where only a subset of system variables are observable.

[3]We use the maximum simulation time T to denote the trajectory length in general. In *Cancer* experiments, the simulation stops if the subject dies intermittently.

[4]Note the difference from some prior preference-based RL works [21], [22], no dedicated human overseer is required in the loop to compare pairwise treatment strategies in our setting. Instead, only the treatment goal composed of one or more evaluation criteria is required before policy learning.

Algorithm 1 PREFERENCE-BASED RL FRAMEWORK

Require:
S': initial states of sampled subjects
N: number of training iterations
T: the maximum simulation time to treat each subject

Randomly initialize $\theta_R, \theta_P^1, \theta_P^2$
$\mathcal{D} = \emptyset, \Gamma^1 = \emptyset, \Gamma^2 = \emptyset$
for $n = 0$ **to** $N-1$ **do** // One training iteration
  for all $s \in S'$ **do** // One sampled subject
    for $i \in [1, 2]$ **do** // One agent
      $s_0^i \leftarrow s, \tau^i \leftarrow \emptyset$
      for $t = 0$ **to** $T-1$ **do**
        $a_t^i \leftarrow \pi(s_t^i; \theta_P^i)$
        $s_{t+1}^i \leftarrow \textsc{Simulate}(s_t^i, a_t^i)$
        $r_t^i \leftarrow \textsc{Reward}(s_t^i, a_t^i; \theta_R)$
        $\tau^i \leftarrow \tau^i \cup \{(s_t^i, a_t^i, r_t^i)\}$
      end for
      $\Gamma^i \leftarrow \Gamma^i \cup \{\tau^i\}$
    end for
    $pre(\tau^1, \tau^2) \leftarrow \textsc{Evaluate}(\tau^1, \tau^2)$ // Follow Treatment goal
    $\mathcal{D} \leftarrow \mathcal{D} \cup \big(\tau^1, \tau^2, pre(\tau^1, \tau^2)\big)$
  end for
  Drawing minibatches $\Gamma_n^1 \sim \Gamma^1, \Gamma_n^2 \sim \Gamma^2, \mathcal{D}_n \sim \mathcal{D}$
  Fitting the reward function θ_R with $\mathcal{D}_n$
  Optimizing the policy θ_P^1 with Γ_n^1, θ_P^2 with Γ_n^2
end for

A. Fitting the Reward Function

Our reward model is parameterized by θ_R with a deep neural network, which takes the state-action pair (s_t, a_t) as input and produces an estimated reward $r_t \in \mathbb{R}$. To train the reward model with preference between two given trajectories, we follow the Bradley–Terry model [24] and adopt the common practice in existing preference-based RL work [16], [21]–[23], where the reward model is interpreted as a preference predictor. Specifically, it is assumed that the probability of preferring trajectory τ depends exponentially on the value of the discounted reward sum over the length of the trajectory:

$$p(pre(\tau^1, \tau^2) = \tau^1 \succ \tau^2; \theta_R) = \frac{\exp R(\pi^1, s_0; \theta_R)}{\exp R(\pi^1, s_0; \theta_R) + \exp R(\pi^2, s_0; \theta_R)}, \tag{1}$$

where capital R denotes the expected return following policy π for a patient with initial state s_0.

We then cast the reward learning problem as a classic binary classification task, where two trajectories are given and the reward model learns to approximate the preference between the two. Therefore the cross-entropy loss between the predictions and the actual label determined by the *treatment goal* is minimized:

$$L(\theta_R) = -\sum_{(\tau^1, \tau^2, pre(\tau^1, \tau^2)) \sim \mathcal{D}_n} \Big(I(\tau^1 \succ \tau^2) \log p(\tau^1 \succ \tau^2; \theta_R) + I(\tau^2 \succ \tau^1) \log p(\tau^2 \succ \tau^1; \theta_R) \Big), \tag{2}$$

where $I(\cdot \succ \cdot)$ is an indicator function determined by $pre(\cdot, \cdot)$, which equals 1 if the first is preferred to the second, 0 otherwise.

B. Optimizing the Policy

At timestep t, the agent observes state s_t, takes action a_t, and receives $r_{\theta_R}(s_t, a_t)$ from the reward model. We propose to use the following two kinds of reward definitions for policy learning:

- **Action-based Reward Modification (*AbRM*)**: the preference-based reward $r_{\theta_R}(s_t, a_t)$ is directly utilized by the agent for policy learning.
- **State-based Reward Modification (*SbRM*)**: We derive a new state value h_{θ_R} from r_{θ_R} to represent how good the current state is: $h_{\theta_R}(s_t) = \max_a r_{\theta_R}(s_t, a)$. We then compute the advantage value of the current state over the previous: $h_{\theta_R}(s_t) - h_{\theta_R}(s_{t-1})$, as the final reward for the policy learner.

Since the preference-based reward is a non-stationary value approximated by a neural network, we implement agents with the policy gradient algorithm, which is robust to changes in the reward function [12], [21]. We maximize the expected return by repeatedly estimating the gradient, and optimizing the policy with gradient descent. To avoid high variance in policy updates, we subtract a baseline b (determined by the current state only) from the expected return R [25]:

$$L(\theta_P) = -\sum_{\tau \in \Gamma_n} \Big(\sum_{t=0}^{|\tau|} \log \pi(a_t|s_t) \big(\sum_{t'=t}^{|\tau|} (R - b_{t'}) \big) \Big). \tag{3}$$

IV. EXPERIMENTS

By answering the following research questions about preference-based RL, we focus on building a comprehensive understanding of its effects on treatment recommendation.

- **RQ1-*Utility***: Does the preference-based qualitative feedback really benefit policy learning as opposed to handcrafted rewards or other treatment recommendation approaches?
- **RQ2-*Goal Correlated***: Does the reward function inferred from preference faithfully follow humans' treatment goals?
- **RQ3-*Reward Transferability***: Is it beneficial to transfer the inferred rewards trained for one scenario to other relevant ones?
- **RQ4-*Agent Design***: How to design RL agents so that *treatment goal*-directed preference over their trajectories leads to more accurate reward estimation and more aligned outcomes with the *treatment goal*?

A. Simulation Settings

We first construct a simulation platform with disease evolution and treatment reactions modeled for *Cancer* and *Sepsis*, and then conduct experiments to evaluate treatment outcomes of different recommendation approaches. Please refer to our website [5] for a detailed description of the simulation platform and implementation details.

[5] https://sites.google.com/view/tr-with-prl/simulations

*a) **Dosage Recommendation for Cancer**:* We use the mathematical model proposed by [6] to simulate cancer evolution and drug treatment effects with random state initialization. Each time-step represents one month in the real world. The agent receives the current tumor size $y_t \in \mathbb{R}^+$ and toxicity level $x_t \in \mathbb{R}^+$, and then suggests a dosage amount $d_t \in \{0.1, 0.4, 0.7, 1.0\}$ to the subject. The subject's health condition updates until the end of the 6-month treatment ($T = 7$) or stops if the subject dies intermittently according to the state-based hazard function. We consider three kinds of *treatment goals* during *Cancer* treatment:

- ***CE***: maximize clinical efficacy, i.e., survival rate.
- ***CE&OF-I***: maximize clinical efficacy and mitigate negative effects represented by the sum of the final tumor size and the toxicity level: $y_{T-1} + x_{T-1}$.
- ***CE&OF-II***: similar to *CE&OF-I* with negative effects represented by two separate health signs: highest toxicity level during treatment $\max_{t=0}^{T-1} x_t$, final tumor size y_{T-1}.

*b) **Blood Purification for Sepsis**:* We employ the mathematical model derived by [26] to simulate the acute inflammation process in response to an infection, where parameters for subjects are calibrated so that the generated trajectories without treatment closely follow observed temporal patterns in the real world. Each time-step represents 0.1 hour in the real world. The environment is partially observable, and the agent can observe only 8 out of 19 physiological features that govern *Sepsis* dynamics. At each time-step t between 5^{th} to 18^{th} hour[6], the agent takes an action $a_t \in \{0, 1\}$ to decide whether to perform a 2-hour blood purification operation. After 100-hour simulation ($T = 1000$), the survival status is determined by one of the physiological features. Besides maximizing survival rate (***CE***), we also set another treatment goal as avoiding too frequent operations: $\sum_{t=0}^{T-1} a_t$ (***CE&OF***).

B. Compared Approaches

We compare results from *AbRM* and *SbRM* with the following approaches proposed in treatment recommendation literature:

• Non-learning [6], [17]: 1) *Constant*: A static dosage amount is given to all subjects for six months; 2) *Random*: One of the four dosage options is randomly selected at each time-step; 3) *Upper Bound*: The subjects with *Sepsis* receive operations all the time throughout the simulation period[7].

• Preference Learning [17]: in the Preference-Based Policy Iteration (*PBPI*), one action is preferred to the other based on their outcomes after certain times of simulations.

• Reinforcement Learning with hand-crafted Reward: 1) *Single-objective RL* [27]: the conventional policy gradient approach; it receives $+1$ for survival, -1 for death, and 0 for all intermediate steps. 2) *Single-objective RL (Ensemble)*: two agents are trained independently, and the one with better performance on the validation set is evaluated on the testing set. It is developed for fair comparison as two agents with different parameter initializations are used in the preference-based RL framework. 3) *Existing Multi-objective RL* [6]: manually defined rewards based on key factors are assigned to each time-step. 4) *Grid-search Multi-objective RL* [28]: both survival rate and the negative impacts are treated as objectives and employ the best linear scalarization retrieved from grid search.

[6]We set the operation time between 5-th and 18-th hour to make the task not too easy (earlier treatment mainly results in survival outcomes) or too infeasible (later treatment has little positive effects and subjects can rarely survive).

[7]In *Upper Bound*, receiving 2-hour treatments all the time equals 7 times of operation during $5\sim18$-th hour.

TABLE I: *Cancer* dosage recommendation: Main results under treatment goal of *CE* and *CE&OF-I*. The treatment goal is to maximize *Survival Rate* (*CE*), and mitigate side effects in terms of final tumor size and toxicity level (*CE&OF-I*). The best result per metric is marked in boldface. We present $avg \pm stdev$ values for experiments with 10 random seeds.

Method	Survival Rate	Tumor+Toxicity
Constant Best (0.4)	19.91%±0.58%	2.22±0.04
Constant Worst (0.1)	4.89%±0.68%	3.72±0.03
Random	17.81%±0.91%	2.23±0.04
PBPI	21.79%±0.64%	2.21±0.07
Single-objective RL	26.96%±3.02%	1.16±0.48
Single-objective RL (*Ensemble*)	27.38%±3.32%	1.14±0.49
Existing Multi-objective RL	18.84%±5.77%	2.28±0.66
Grid-search Multi-objective RL	28.98%±3.42%	0.66±0.45
AbRM (*CE*)	31.52%±1.38%	0.46±0.06
AbRM (*CE & OF-I*)	31.33%±1.18%	**0.39±0.02**
SbRM (*CE*)	30.54%±3.46%	0.68±0.45
SbRM (*CE&OF-I*)	**31.72%±1.08%**	0.43±0.06

C. Performance Comparison (RQ1)

*a) **Cancer**:* We list treatment outcomes of different approaches for the first two goals (*CE* and *CE&OF-I*) in Table I and the third goal (*CE&OF-II*) in Table II [8].

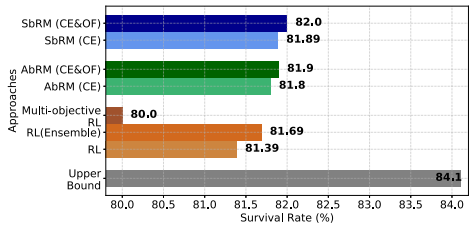
Table I: Considering *Survival Rate* maximization as the only *treatment goal*, agents learning from either action-based (31.52%) or state-based (30.54%) preference reward have much better performance in saving lives than *Single-objective RL* (26.96%), where the hand-crafted reward is used. When negative impacts (*CE&OF-I*: sum of final tumor size and toxicity level) are expected to be mitigated besides saving lives, agents receiving rewards inferred from preference are capable of maintaining the high survival rate with low negative impacts than other baselines.

Table II: When the highest toxicity level during treatment and final tumor size are separate goals to be accomplished besides maximizing survival rate (*CE&OF-II*), we observe that preference-based reward guides the agent to policies with the highest survival rates, while one negative impact gets reduced but the other increases compared with other approaches. This is expected since the two side-goals are conflicting in nature: large amounts of dosage result in higher toxicity levels but smaller tumor sizes and vice-versa.

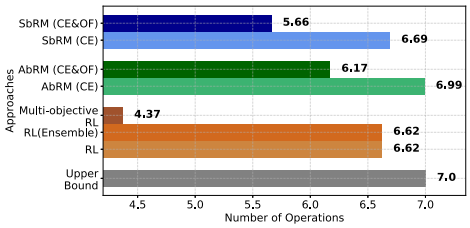
[8]https://sites.google.com/view/tr-with-prl/appendix

Fig. 2: *Sepsis* operation recommendation: (a)(b) Main results under treatment goal of *CE* and *CE&OF*; (c) Evolution of treatment outcome indicator PI on testing septic subjects; (d)(e) Highly-correlated expected return of inferred reward and survival rate; (f)(g) Reward function pre-trained on 2-hour operation task provides better initial performance on 4-hour setting.

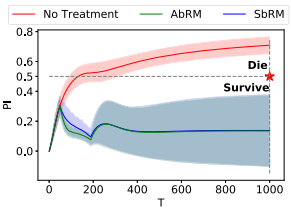
(a) Maximize Survival Rate Goal for *CE* and *CE&OF*



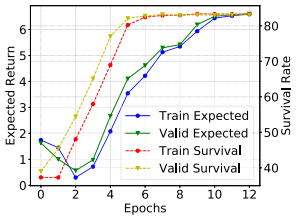
(b) Minimizing #Operations Goal for *CE&OF*



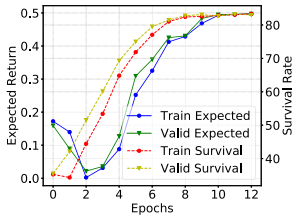
(c) Evolution of PI Value (*CE*)



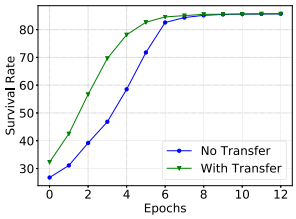
(d) *AbRM* Training (*CE*)



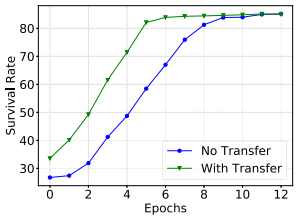
(e) *SbRM* Training (*CE*)



(f) *AbRM* Transferability (2→4h) (*CE*)



(g) *SbRM* Transferability (2→4h) (*CE*)



b) ***Sepsis****:* Figure 2a and 2b illustrate the performance bar charts of different approaches evaluated by *Survival Rate* and *Number of Operations*. When guided by preference-based reward rather than manually defined reward, a slightly higher *Survival Rate* is achieved by both *AbRM* and *SbRM*, while the average number of operations has fallen considerably, by 6.79% with *AbRM* and 14.50% with *SbRM*. Note that although *Multi-objective RL* leads to the fewest number of operations, the resulting *Survival Rate* drops down to make undesired trade-offs between survival rate maximization and negative impact mitigation. In Fig. 2c, we also illustrate the evolution of the treatment outcome indicator PI during $T = 1000$ simulation for $1,000$ testing *Septic* subjects without and with blood purification treatment from preference-based RL agents.

D. Compared with Hand-Crafted Rewards (RQ1)

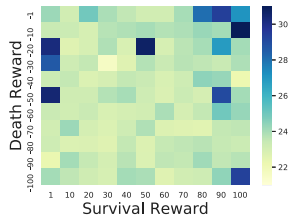
a) ***Optimal Policies with Low Variances****:* As shown in Table I and Table II, compared with conventional *Policy Gradient* with hand-crafted rewards, the RL agents learning with rewards inferred from preference are able to provide the highest survival rate and smallest side effects with relatively small variance under different random seeds. Besides the common practice of assigning $-1/+1$ rewards to trajectories with death/survival outcomes [1]–[3], we here investigate performance of *Policy Gradient* with more comprehensive hand-crafted reward designs (absolute reward value ranges from 1 to 100), and show results in Fig. 3a, 3b and 3c. The different distributions of good policies (darker blue grids concentrated around upper right in Fig. 3a, lower left in Fig. 3b and lower right in Fig. 3c) in three random seeds illustrate the difficulty of finding a stationary reward design with consistently good performance. When learning from hand-crafted rewards under distinct random seeds, the *Policy Gradient* approach converges to different local minima, resulting in average sub-optimal performance with high variance.

b) ***Priorities Recognized among Different Treatment Goals****:* In Fig. 3d, 3e and 3f, we further show suboptimal performance obtained by *Multi-objective RL* with grid-searched linear scalarization, when humans aim to realize multiple *treatment goals*, some of which are competing with each other (e.g., more dosage treatment causes smaller tumor but higher toxicity). Different ratios of objective factors used in the reward function cause negative interference between minimizing *tumor size* and reducing *toxicity level*, while the RL agent can hardly recognize the priority of saving subjects' lives over mitigating the other two side effects. In the end, the *Multi-objective RL* approach achieves quite low survival rate, although with small tumor size (red zones in Fig. 3d or 3e) or low toxicity level (blue zones in Fig. 3d and 3f). Instead of using a linear combination of goal factors as rewards, the proposed framework infers the reward function with *treatment goal*-directed preference and is able to prioritize maximizing survival rate over other goals, which aligns well with human intent.

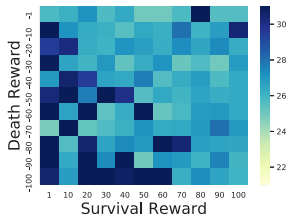
c) ***Incorporating Incomparable Trajectories for Better Performance****:* Given two treatment trajectories for one sampled subject, if they have identical performance according to human's objectives, then the two trajectories are deemed to be incomparable. Take *Patient C* in Fig. 1a as an example: neither of the two policies τ_C^1 and τ_C^2, should be preferred since they both results in deaths. Therefore this trajectory pair is incomparable, i.e., $\tau_C^1 \sim \tau_C^2$. Since no clear preference conclusion can be drawn between the two incomparable trajectories, the majority of existing work in preference learning disregarded them directly [17]–[21]. Only comparable pairs, either τ^1 preferred to τ^2 ($I(\tau_C^1, \tau_C^2) = 1$) or τ^2 preferred to τ^1

Fig. 3: *Cancer* dosage recommendation: (a)(b)(c) Different distributions of good policies learned by conventional *Policy Gradient* from diverse reward designs for survival/death outcomes; (d)(e)(f) *Multi-objective RL* learning from linearly weighted reward function leads to sub-optimal performance: low *survival rate* although with small tumors (red zones) or low toxicity (blue zone); the linear weight assigned to each factor is one of four values: $\{1, 2, 4, 8\}$ and each marker represents performance obtained by one of the 37 combinations.

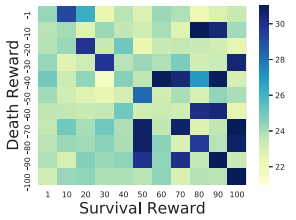
(a) *PG* for *CE*: Seed 2001



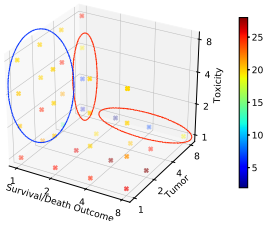
(b) *PG* for *CE*: Seed 2002



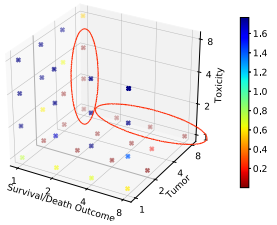
(c) *PG* for *CE*: Seed 2003



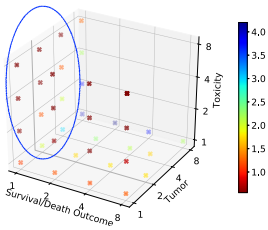
(d) MORL for *CE&OF-II*: Survivals



(e) MORL for *CE&OF-II*: Tumor



(f) MORL for *CE&OF-II*: Toxicity



($I(\tau_C^2, \tau_C^1) = 1$), are included in the training set to optimize preference approximation. However, preference learning based on comparable trajectories alone achieves quite unsatisfactory *survival rate* in our treatment recommendation tasks (green curve in Fig. 4c). Two reasons are likely to contribute to this failure: 1) polarized preference (one preferred with probability 0.75, and the other 0.25 is inferred in Fig. 4a) between two incomparable trajectories although the preference label is never provided in the training set; 2) only around one-fifth of the trajectory pairs (2,000 comparable from 10,000 sampled subjects) are leveraged in each epoch for reward model update (green line in Fig. 4b). After the above performance analysis, we find that excluding incomparable pairs from the training set leaves the parameterized reward model exploring the reward space arbitrarily and inferring random preferences between two trajectories even when they are incomparable. To avoid arbitrary exploration in the reward space, we handle the incomparable pairs with a simple approach: treating both trajectories from the incomparable pair equally, i.e., $I(\tau^1 \succ \tau^2) = I(\tau^2 \succ \tau^1) = 0.5$. With this small but important augmentation to the preference indicator function, incomparable trajectory pairs are efficiently utilized for: 1) better reward space exploration: preference approaching 0.5 as expected in Fig. 4a; 2) more samples utilized for reward model update: all the $10,000$ samples from the training set participate in the loss function minimization in Fig. 4b, and 3) much better treatment outcomes: more than 30% survival rate achieved after the model converges in Fig. 4c[9].

E. Correlation with Treatment Goals (RQ2)

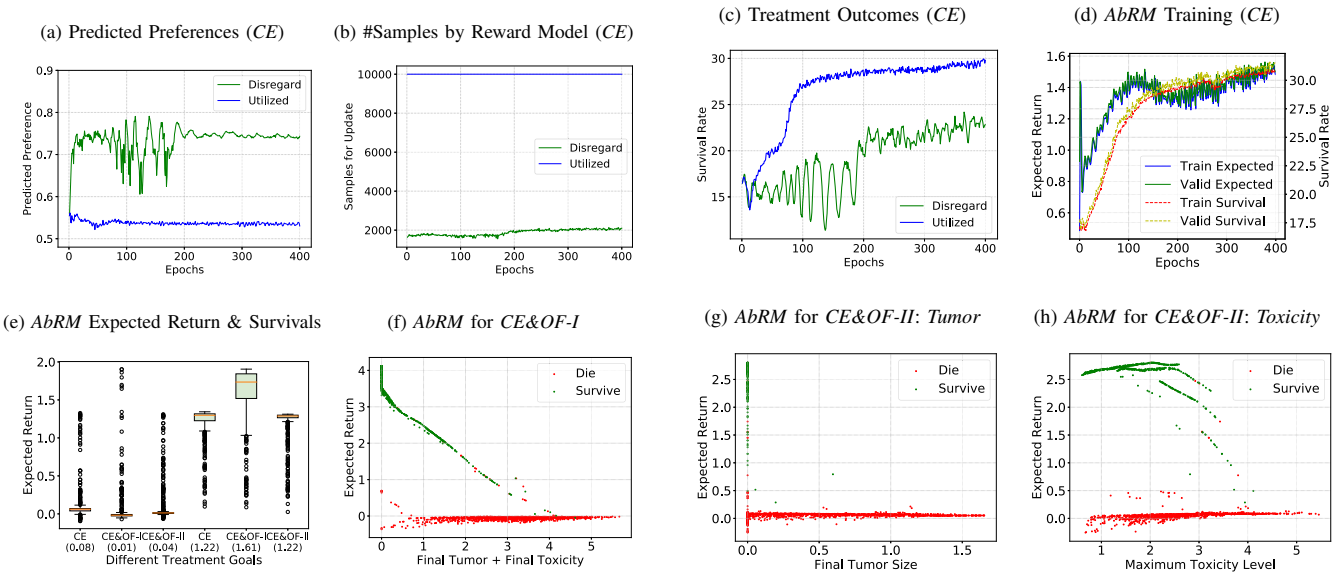
Although resulting in suboptimal treatment outcomes, handcrafted rewards are closely connected with the key factors that the *treatment goal* has considered. For instance, $-1/+1$ assigned to trajectories with death/survival outcomes for goal *CE*, a linear weighted sum of survival status, tumor and toxicity for goal *CE&OF-II*, etc. On contrary, whether the inferred rewards in preference-based RL are correlated to human's *treatment goal* is unclear, since the two are indirectly connected by preference. Therefore, we study how well the preference-based reward matches humans' actual treatment intentions by visualizing the expected return of inferred rewards and important factors from *treatment goals*.

a) ***Training****:* During training *AbRM* for *Cancer* experiments, we can observe from Fig. 4d, that the rising trend of the expected return matches the improving *Survival Rate* quite well, although the parameters of the reward and policy model are being updated simultaneously. For *Sepsis* experiments, we also observe a positive correlation between the expected return and survival rate in Fig. 2d and 2e.

b) ***Testing****:* After the model converges, we demonstrate the correlation between expected return of inferred rewards and treatment outcomes in Fig. 4e. Under different *treatment goals*, the reward model is able to prioritize the treatment outcome over other side effects, so that treatment trajectories leading to death outcomes always have extremely low expected return (approaching 0) while those saving subjects' lives usually get much higher expected return. The expected return for trajectories with survival outcomes further shows their correlation with side effect mitigation goals in Fig. 4f, 4g and 4h. The reward function can not only distinguish trajectories leading to different survival statuses (green dots are far away from red dots), but also differentiate trajectories with same survival outcomes but different negative impacts: 1) the expected return is negatively proportional to the sum of tumor and toxicity for goal *CE&OF-I* in Fig. 4f; 2) when the two goals are competing with each other for goal *CE&OF-II*, we can still observe clear decreasing trend from expected return when toxicity level increases in Fig. 4h. Based on the above analysis, we conclude that the rewards inferred from preference are highly

[9]All the reported results of preference-based RL models *AbRM* and *SbRM* in this paper take incomparable trajectories into consideration as opposed to previous works.

Fig. 4: *Cancer* dosage recommendation: (a)(b)(c) Comparison between utilizing and disregarding incomparable trajectories to infer rewards; (d): Highly-correlated expected return of inferred reward and survival rate during training; (e)(f)(g)(h): Correlation between expected return of inferred reward and different *treatment goals* during testing.



correlated to different *treatment goals*, and can prioritize the most important one (i.e., maximizing *survival rate*) from others (i.e., mitigating diverse side effects), which is lacking in hand-engineered rewards. Similar conclusions can be obtained from model *SbRM* (without visualizations here due to space constraints).

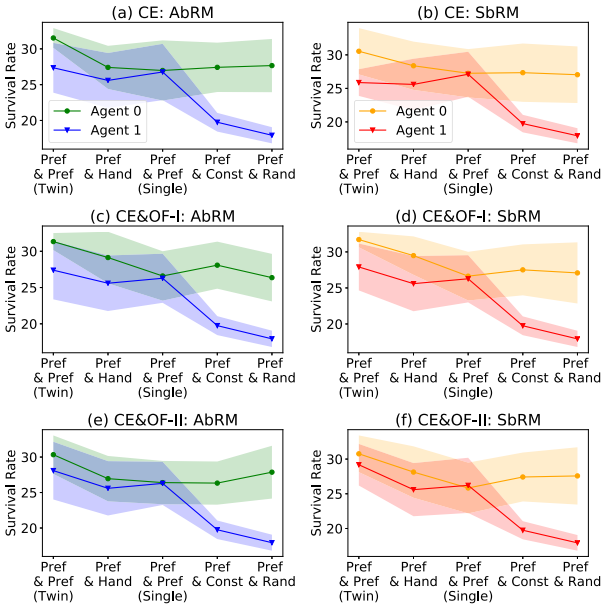
F. Reward Transferability (RQ3)

According to Eq. 1, the probability of preferring trajectory τ depends exponentially on the value of the discounted sum of the reward over the length of the trajectory. With *treatment goal* and trajectory length fixed, a well-trained reward model is supposed to keep its capability of distinguishing the trajectory more aligned with human's intention from the other, when some experiment settings vary in another scenario. We therefore investigate the transferability capability of preference-based rewards by extending every blood purification operation for septic subjects to 4 hours and observing the performance of a reward model which has been pre-trained on the 2-hour operation setting. As expected in Fig. 2f and 2g, the pre-trained reward function provides higher initial survival rate compared with learning rewards from scratch, which proves its transferability in assigning higher rewards to preferred treatment trajectories in similar but different application scenarios. This signifies a potential for faster convergence and reduced training cost when the proposed preference-based RL framework is adapted from simulation to the real world.

G. Agent Design Guidelines (RQ4)

As introduced in Algorithm 1, the proposed preference-based RL framework adopts two RL agents controlled by different parameters to learn the preference-based reward. We therefore study the influence of different agent designs on

Fig. 5: *Cancer* dosage recommendation: Effects of distinct agent designs for reward inference on performance with different treatment goals. Agents learning from preference-based reward is denoted by *Pref* and hand-crafted reward by *Hand*. We use *Twin* to indicate two agents with different parameters while *Single* for shared parameters. *Rand* and *Const* represent two baselines: *Random* and *Constant Best*.



reward approximation and the resulting performance. Specifically, the reward function is estimated to approximate the preference over trajectories, among which the first trajectory is performed by one RL agent learning from preference-

based reward, while the second trajectory can be executed by different agent types. First of all, the survival rate curves shown in Fig. 5 empirically prove the effectiveness of the current design of two different RL agents both learning from preference-based rewards. Based on the descending order of performance from both *Agent 0* and *Agent 1* in Fig. 5, we can observe that the performance ranking of agents (*Agent 0*) learning from preference-based reward: *Pref&Pref(Twin) > Pref&Hand > Pref&Const > Pref&Rand*, is consistent with their competitor's (*Agent 1*) ranking listed in Table I: *AbRM/SbRM > Single-objective RL > Constant Best (0.4) > Random*. We therefore attribute the success of the preference-based RL framework to the preference over trajectories from equally competitive agents, which leads to more accurate reward estimation, better and more competitive policies from both agents, and continues till both agent and reward models converge. Besides, we also notice a drop in performance when two preference-based RL agents share the same network. Since the diversity in trajectories now mainly comes from stochasticity in the simulation platform, we suspect that the performance suffers due to a lack of adequate exploration.

V. Conclusion

To obtain optimal treatment policies based on human's diverse *treatment goals*, we investigate the performance of the preference-based reinforcement learning approaches. Specifically, the preference over two treatment trajectories for one sampled subject is evaluated according to the human's *treatment goal* and the reward is estimated based on the preference. With the constructed simulation platform, we systematically examine the preference-based RL framework and observe its high performance and close correlation between inferred rewards and *treatment goals*. Further, we explore reward transferability and present an agent design study for a deeper understanding in designing preference-based RL approaches and to better aid clinicians with useful treatment strategies.

References

[1] A. Raghu, M. Komorowski, I. Ahmed, L. Celi, P. Szolovits, and M. Ghassemi, "Deep reinforcement learning for sepsis treatment," *arXiv preprint arXiv:1711.09602*, 2017.

[2] L. Wang, W. Zhang, X. He, and H. Zha, "Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 2447–2456.

[3] S. Nemati, M. M. Ghassemi, and G. D. Clifford, "Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 2978–2981.

[4] B. T. Denton, M. Kurt, N. D. Shah, S. C. Bryant, and S. A. Smith, "Optimizing the start time of statin therapy for patients with diabetes," *Medical Decision Making*, vol. 29, no. 3, pp. 351–367, 2009.

[5] D. Lopez-Martinez, P. Eschenfeldt, S. Ostvar, M. Ingram, C. Hur, and R. Picard, "Deep reinforcement learning for optimal critical care pain management with morphine using dueling double-deep q networks," *arXiv preprint arXiv:1904.11115*, 2019.

[6] Y. Zhao, M. R. Kosorok, and D. Zeng, "Reinforcement learning design for cancer clinical trials," *Statistics in medicine*, vol. 28, no. 26, pp. 3294–3315, 2009.

[7] C. Wirth, R. Akrour, G. Neumann, and J. Fürnkranz, "A survey of preference-based reinforcement learning methods," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4945–4990, 2017.

[8] D. M. Faissol, P. M. Griffin, and J. L. Swann, "Timing of testing and treatment of hepatitis c and other diseases," in *Proceedings*, 2007, p. 11.

[9] T.-H. Pham, G. De Magistris, and R. Tachibana, "Optlayer-practical constrained optimization for deep reinforcement learning in the real world," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 6236–6243.

[10] T. Schaul, D. Borsa, J. Modayil, and R. Pascanu, "Ray interference: a source of plateaus in deep reinforcement learning," *arXiv preprint arXiv:1904.11455*, 2019.

[11] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 1.

[12] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Advances in neural information processing systems*, 2016, pp. 4565–4573.

[13] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care," *Nature Medicine*, vol. 24, no. 11, p. 1716, 2018.

[14] S. Saria, "Individualized sepsis treatment using reinforcement learning," *Nature medicine*, vol. 24, no. 11, pp. 1641–1642, 2018.

[15] Y. Gao, H. Xu, J. Lin, F. Yu, S. Levine, and T. Darrell, "Reinforcement learning from imperfect demonstrations," *arXiv preprint arXiv:1802.05313*, 2018.

[16] D. S. Brown, W. Goo, P. Nagarajan, and S. Niekum, "Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations," *arXiv preprint arXiv:1904.06387*, 2019.

[17] J. Fürnkranz, E. Hüllermeier, W. Cheng, and S.-H. Park, "Preference-based reinforcement learning: a formal framework and a policy iteration algorithm," *Machine learning*, vol. 89, no. 1-2, pp. 123–156, 2012.

[18] W. Cheng, J. Fürnkranz, E. Hüllermeier, and S.-H. Park, "Preference-based policy iteration: Leveraging preference learning for reinforcement learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2011, pp. 312–327.

[19] R. Akrour, M. Schoenauer, and M. Sebag, "April: Active preference learning-based reinforcement learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, pp. 116–131.

[20] D. Schäfer and E. Hüllermeier, "Preference-based reinforcement learning using dyad ranking," in *International Conference on Discovery Science*. Springer, 2018, pp. 161–175.

[21] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems*, 2017, pp. 4299–4307.

[22] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei, "Reward learning from human preferences and demonstrations in atari," *arXiv preprint arXiv:1811.06521*, 2018.

[23] D. S. Brown, W. Goo, and S. Niekum, "Better-than-demonstrator imitation learning via automatically-ranked demonstrations," in *Conference on Robot Learning*. PMLR, 2020, pp. 330–359.

[24] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[25] J. Schulman, N. Heess, T. Weber, and P. Abbeel, "Gradient estimation using stochastic computation graphs," *arXiv preprint arXiv:1506.05254*, 2015.

[26] S. O. Song, J. Hogg, Z.-Y. Peng, R. Parker, J. A. Kellum, and G. Clermont, "Ensemble models of neutrophil trafficking in severe sepsis," *PLoS computational biology*, vol. 8, no. 3, 2012.

[27] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2015.

[28] I. Y. Kim and O. De Weck, "Adaptive weighted sum method for multiobjective optimization: a new method for pareto front generation," *Structural and multidisciplinary optimization*, vol. 31, no. 2, pp. 105–116, 2006.